UNIVERSITÀ DEGLI STUDI DI PADOVA

Dipartimento di Fisica e Astronomia “Galileo Galilei”

Corso di Laurea in Fisica

Tesi di Laurea

Analisi statistica delle proprietà di complessità topologica in strutture di proteine

Relatore:

Prof. Antonio Trovato

Laureando:

Edoardo Antonaci

Anno Accademico 2018/2019

Capitolo 1: Una panoramica sulle proteine

1.1 Struttura delle proteine.

1.2 Ripiegamento delle proteine.

1.3 Complessità topologica nelle proteine.

Capitolo 2: Analisi topologica

2.1 Procedura di clustering.

2.2 Asimmetria e chiralità.

2.3 Distribuzione delle lunghezze dei loop.

Capitolo 3: Conclusioni

Discussione dei risultati raggiunti e prospettive future

Capitolo 1

UNA PANORAMICA SULLE PROTEINE

1.1 **Struttura delle proteine**

Le proteine sono grandi complessi molecolari polimerici formati da unità di base chiamate amminoacidi e svolgono funzioni essenziali per la vita della cellula. Il loro principale scopo è quello relativo alla catalisi enzimatica legato alle conversioni chimiche all'interno e all'esterno della cellula.

Esistono, per esempio, proteine regolatrici che controllano l'espressione genetica e proteine recettrici, situate all'interno della membrana lipidica, che accettano segnali intracellulari che sono spesso trasmessi da ormoni (proteine a loro volta). Le immunoproteine, invece, riconoscono e legano molecole "nemiche" così come cellule "amiche", aiutando le seconde ad essere più facilmente inglobate nell'organismo.

Sono sempre proteine le molecole responsabili del trasferimento transmembranico di protoni ed elettroni, necessario per interi cicli bioenergetici come, per esempio: l'assorbimento di luce, la respirazione e la produzione di ATP. Quest'ultimo permette attività chimiche e meccaniche all'interno del nostro corpo come la contrazione dei muscoli e il movimento di molecole nella cellula.

L'enorme varietà delle funzioni proteiche è dovuta all'alta specificità mostrata verso le molecole con le quali interagiscono, una relazione che ricorda quella della chiave e della serratura. Questa specifica relazione, richiede una struttura spaziale precisa della proteina. Questo è il motivo per cui le funzioni biologiche delle proteine (così come di altre macro molecole di importanza fondamentale per la vita come il DNA e RNA) sono così connesse con la loro struttura tridimensionale.

Anche un piccolo danneggiamento di queste strutture è spesso la ragione per la perdita o il cambiamento drammatico dell'attività proteica. Una conoscenza della struttura 3D di una proteina è, quindi, necessaria per capire come essa funzioni [4].

Diversi sono i metodi che possono essere utilizzati per la determinazione, in toto o in parte, della struttura proteica. Semplificando, possiamo suddividerli in due distinte classi metodologiche: una che include procedimenti che coinvolgono diffrazione o scattering sia di particelle subatomiche che di onde elettromagnetiche; un'altra che, invece, include metodi spettroscopici che sfruttano i cambiamenti degli stati energetici degli atomi della proteina. Questi hanno luogo come risultato delle interazioni di atomi con la radiazione elettromagnetica di differenti frequenze.

Un esempio della prima categoria è la diffrazione a raggi X. Come è noto, il potere risolutivo dell'occhio umano si aggira intorno ai 0.2mm; questo vuol dire che due punti che distano meno di questa soglia vengono interpretati dal nostro cervello come un unico punto. Nelle molecole delle proteine gli atomi si trovano a distanza molto più piccola (circa 1 o 2 $\dot{A}$), rendendo quindi impossibile distinguere ad occhio nudo simili strutture. Per questo si usano i raggi X che avendo una lunghezza d'onda tra i 0.1 e 100 $\dot{A}$, riescono a "vedere" gli atomi.

C'è però un problema dovuto al fatto che non può esistere un "microscopio a raggi X" perché questo tipo di onde, differentemente dalla luce visibile, non può essere focalizzato dalle lenti. Per ovviare a questa mancanza si è pensato ad un metodo indiretto: guardare come le molecole diffrangono i raggi X.

Conoscendo la posizione degli atomi è poi possibile, tramite algoritmi matematici, ricostruire l'intera struttura proteica. In realtà vi è un passaggio intermedio a questo processo: la cristallografia. In pratica, a causa della presenza di atomi molto leggeri (quindi con pochi elettroni come C, N e O) nell'"ossatura" delle proteine, i raggi X vengono diffratti poco e occorre quindi direzionare il più possibile le molecole nello spazio in modo da creare un "cristallo" e avere scattering più marcati.

In alternativa, come esempio della seconda categoria, si può usare la risonanza magnetica nucleare. In questo caso si prende in considerazione il momento magnetico di spin dei diversi nuclei che costituiscono la proteina (^{1}H, ^{13}C, ^{15}N, ^{19}F; ^{31}P hanno tutti spin pari ad ½); questi quando vengono immersi in un campo magnetico intenso possono o allineare il vettore momento al campo magnetico (facendo diminuire l'energia del sistema) o contrapporlo (facendo quindi aumentare l'energia del sistema). Con opportuni impulsi energetici è possibile far cambiare configurazione a questi atomi che si eccitano e, a loro volta, emettono ben precise lunghezze d'onda che risultano visibili alle apparecchiature come picchi ben definiti.

Inoltre, in base alla posizione dell'atomo, se in corrispondenza del gruppo amminico o del gruppo carbossilico (in un amminoacido -NH_2 è il gruppo amminico mentre -COOH è il gruppo carbossilico) questi segnali assumono forme differenti ed è quindi possibile capire come sono disposti gli atomi e dunque la struttura proteica [2].

Per comprendere la complessità topologica delle proteine è necessario partire dai loro costituenti elementari, ovvero gli amminoacidi.

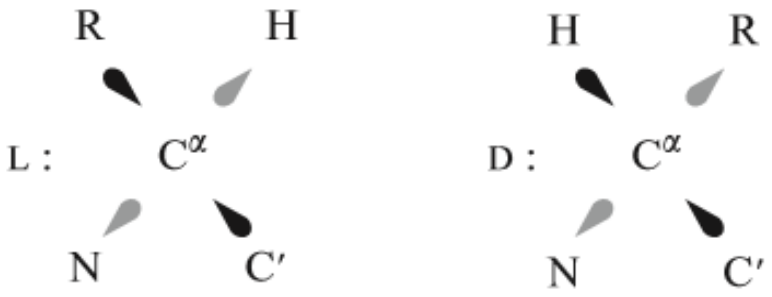


Figura1: forme steriche L e D di un amminoacido

Nella figura precedente si vedono le due forme steriche possibili di un amminoacido: la forma L e la forma D che, come si può notare, non sono simmetriche tra loro.

Sebbene siano state presentate entrambe le forme, solo i residui che provengono dalla forma L permettono alla catena proteica di essere costruita. Solo questi, infatti, sono codificati dai geni. Gli amminoacidi nella forma D non sono tradotti nella matrice proteica, ma sintetizzati solo in appositi enzimi e la loro eliminazione è molto laboriosa.

Con R è stata indicata la catena laterale. Amminoacidi di diverso tipo sono caratterizzati da diverse catene laterali.

Esistono 20 tipi di amminoacidi diversi che vengono uniti gli uni con gli altri grazie al legame peptidico: sostanzialmente, si legano tra loro gli atomi C′ e N di due amminoacidi diversi, liberando una molecola d'acqua, e questo conferisce alla struttura la giusta rigidità e linearità. Questo si spiega considerando che è presente la cosiddetta "ibridizzazione" sp^2: una proprietà puramente quantistica che consente ai due atomi di "condividere" gli elettroni di legame che risultano essere non più "localizzati" su di un unico atomo.

La porzione di amminoacido che rimane dopo la formazione del legame peptidico è chiamata residuo.

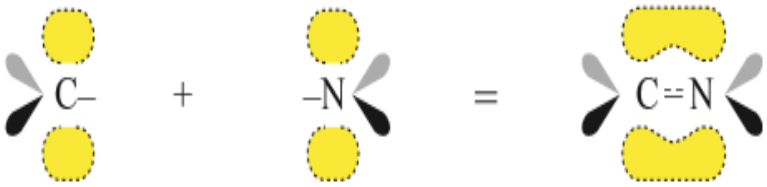


Figura 2: rappresentazione del legame peptidico

Più legami peptidici formano quella che viene definita “schiena” o “catena principale” della proteina dalla quale emergono le catene laterali.

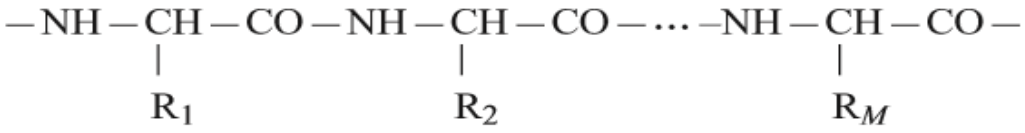


Figura 3: catena peptidica principale

Il numero M dei residui può andare dalle poche dozzine alle diverse migliaia; a seconda dell’informazione codificata nei geni. Come è anche mostrato in figura 3, ha senso parlare di residuo N-terminale e di residuo C-terminale dove si fa riferimento all’azoto del gruppo amminico -NH_2 (primo residuo della catena proteica) e al carbonio del gruppo carbossilico -COOH (ultimo residuo della catena proteica) [5].

1.2 **Ripiegamento delle proteine**

A causa dei gruppi idrofili e idrofobici che formano la catena proteica, la forma tridimensionale della proteina, ovvero quella che siamo soliti vedere in laboratorio, non è liscia e regolare come si può pensare considerando il solo legame peptidico (derivante dalla ibridizzazione sp^2), ma piuttosto complessa e articolata su più livelli strutturali.

Già dalla sua sintesi all’interno del ribosoma (organulo presente all’interno della cellula), la proteina va incontro al cosiddetto “folding cotraslazionale”: mano a mano che gli amminoacidi si uniscono alla catena peptidica tendono a “foldare”, ovvero a ripiegare su stessi mentre la catena viene fatta traslare lungo il tunnel ribosomiale. Si tratta di un processo complesso, che è stato studiato per esempio negli articoli ([8],[9]). Inoltre, una volta fuori dal ribosoma le proteine vanno incontro ad ulteriori riorganizzazioni che vanno sotto il nome di strutture secondarie: alpha elica e foglietto beta pieghettato.

Come si evince dallo stesso nome, l'alpha elica è prettamente un avvolgimento destrorso stabilizzato da legami idrogeno con un passo di 3.6 residui per giro.

Il foglietto beta, invece, è l'insieme di più filamenti peptidici disposti l'uno accanto all'altro e collegati tra loro da due o più legami idrogeno che formano una struttura planare molto compatta.

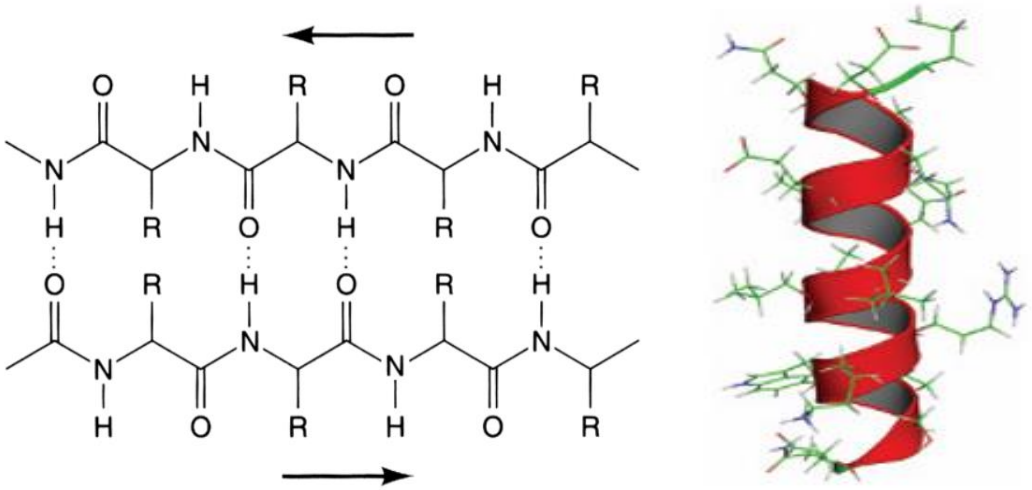
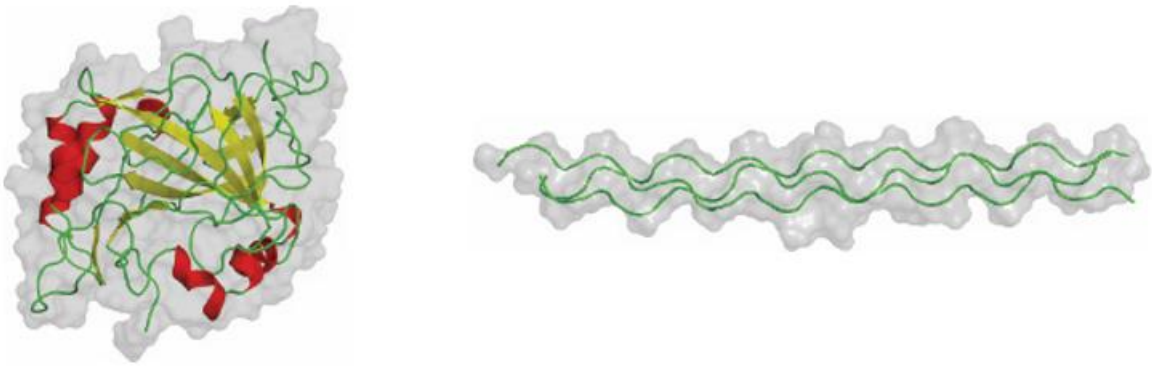


Figura 4: a sinistra vi è un'immagine, presa da https://it.wikipedia.org/wiki/Foglietto_β, di due filamenti del foglietto B antiparalleli , a destra invece è mostrata l'alpha elica.

Mentre la struttura secondaria può essere presente in entrambe le sue forme nella stessa proteina, per specifiche funzioni occorre che la proteina assuma un ulteriore grado di organizzazione che si ottiene ripiegando ulteriormente la catena peptidica dando vita alla cosiddetta struttura terziaria. Essa assume principalmente due forme:

1.*Globulare*: in queste proteine la catena polipeptidica cambia spesso direzione, creando così una struttura nel complesso sferica. La maggior parte delle proteine cellulari sono globulari e svolgono diverse funzioni. La maggior parte delle proteine sono anche solubili in acqua, risiedendo nel citosol o in fluidi di corpi multicellulari. Alcune inoltre risiedono nella membrana plasmatica o (negli eucarioti) nelle membrane intracellulari.

2.*Fibrosa*: la catena polipeptidica di una proteina fibrosa crea una forma allungata e insolubile in acqua. Le proteine fibrose tipicamente svolgono funzioni più semplici se comparate con la controparte globulare. Queste funzioni includono la costruzione di grandi strutture cellulari ed extracellulari, le quali forniscono il supporto meccanico alle cellule e ai tessuti, protezione fisica, o altre funzioni tipiche di specifici tessuti.



Globulare Fibrosa

Figura 5: le due possibili forme per la struttura terziaria

Molte proteine, probabilmente la maggior parte, contengono più catene polipeptidiche che interagiscono le une con le altre, sebbene non covalentemente. L'arrangiamento spaziale di queste catene è indicato con "struttura quaternaria" e le interazioni tra loro sono chiamate "interazioni quaternarie". Le catene sono spesso chiamate subunità. Il termine dominio si usa invece per indicare regioni diverse della stessa catena polipeptidica che ripiegano in maniera autonoma in strutture compatte indipendenti fra loro.

Le strutture quaternarie sono stabilizzate principalmente da interazioni non-covalenti, che permettono l'equilibrio dinamico tra le diverse subunità. Come nel caso del ripiegamento proteico, la formazione della struttura quaternaria è guidata da effetti idrofobici e resa specifica da interazioni elettrostatiche. Queste ultime tendono a legare le subunità facendo spesso da "ponte" tra esse.

Ad oggi il processo di ripiegamento della proteina dalla sequenza amminoacidica lineare fino alla struttura quaternaria è spiegato usando l'ipotesi di Anfinsen: la proteina, che parte come filamento semplice (con massima entropia), assume sempre la conformazione a cui corrisponde il minimo assoluto di energia libera del sistema proteina più solvente. Inoltre Anfinsen ha dimostrato sperimentalmente che tale ripiegamento non è guidato da meccanismi interni alla cellula ma è univocamente determinato dalla struttura primaria, cioè dagli amminoacidi, che compongono la catena peptidica.La proteina ripiega quindi nello stato nativo velocemente ed innumerevoli volte sempre nello stesso modo e questo avviene, sorprendentemente, anche per strutture caratterizzate da topologie molto complesse.

A livello teorico, si ritiene che questa proprietà delle proteine sia il risultato dell'evoluzione naturale, che ha selezionato sequenze di amminoacidi caratterizzate da poca frustrazione e quindi da un caratteristico paesaggio di energia ad imbuto ("funnel landscape") mostrato in figura 6.

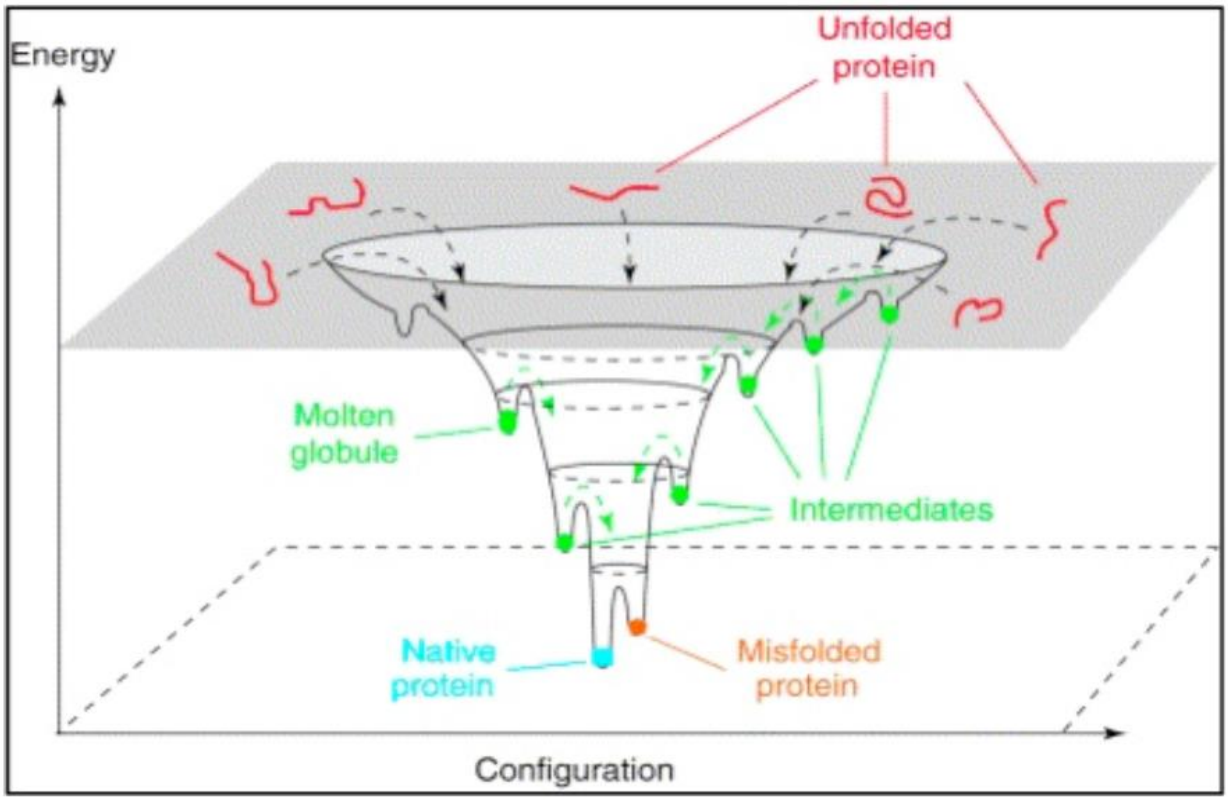


Figura 6 : paesaggio di energia ad imbuto,immagine presa dal sito http://www.federica.unina.it/scienze-biotecnologiche/biochimica-scb/organizzazione-strutturale-proteine/

1.3 **Complessità topologica nelle proteine**

È interessante capire, in particolare, come avviene il ripiegamento cotraslazionale. Quando la proteina, a mano a mano che viene assemblata, è immessa nel tunnel ribosomiale (e in seguito nell'ambiente cellulare) può presentare (e in genere presenta) due o più punti della stessa catena peptidica che interagiscono tra loro, dando vita a un vero e proprio cappio che in seguito sarà chiamato loop.

Le modalità dell'interazione possono essere di varia natura: tramite ponti disolfuro, legami ad idrogeno oppure interazioni di Wan der Waals.

Formato questo loop è importante capire se esso venga attraversato ("threaded") da un altro filamento della catena e se questo filamento sia più vicino al C-terminale o al N-terminale della proteina rispetto al loop. I due casi verranno chiamati rispettivamente C-thread ed N-thread (vedere figura 8).

Approfondire questo aspetto può essere utile per comprendere qualcosa in più della "cinetica di ripiegamento" della proteina che, almeno in teoria, potrebbe dipendere dalle dimensioni di questa e dalla lunghezza del loop, nonché dalla separazione in sequenza tra il filamento e il loop stesso.

La situazione è molto complicata perché, come è stato detto in precedenza, le interazioni tra gruppi idrofili ed idrofobici sono tante e i ripiegamenti che ne seguono portano a strutture proteiche anche molto complesse, come mostrato in figura 7.

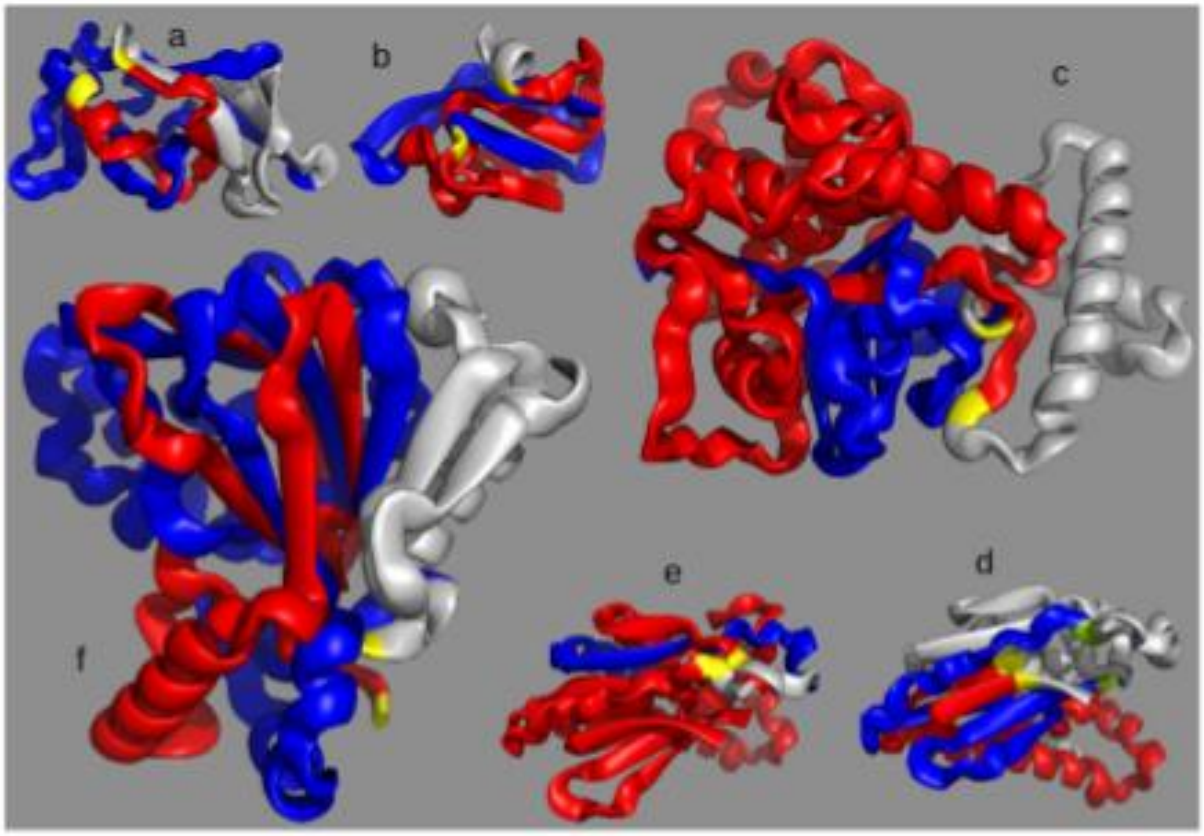


Figura 7: Il filamento blu è il thread (N o C) che si avvolge intorno al loop in rosso delimitato dai due amminoacidi in giallo. Può capitare che il filamento formi a sua volta un cappio intorno al loop "principale" come si vede guardando l'immagine d.

Per tentare uno studio di questo tipo è chiaramente necessario fissare una distanza di soglia affinché si possa dire che due amminoacidi distanti lungo la catena siano vicini tra loro nello spazio tridimensionale. Questa soglia è stata scelta [1] pari a 4.5 $\dot{A}$ e confrontata con le distanze fra atomi "pesanti", cioè ad esclusione dell'idrogeno, presi da due amminoacidi diversi. Se la distanza è minore di 4.5 $\dot{A}$ per almeno una coppia di atomi "pesanti" si può dire che i due amminoacidi sono in contatto e che quindi definiscono le estremità di un loop. In [1] è stato utilizzato un indicatore chiamato "Gaussian entanglement" (GE) definito come

il doppio integrale circolare su una curva chiusa e una aperta: il loop e il filamento appunto. Il GE generalizza la nozione di "linking number" fra due curve chiuse, la cui definizione matematica è sul continuo:

$$G \equiv \frac{1}{4\pi} \oint_{\gamma_i} \oint_{\gamma_j} \frac{\boldsymbol{r}^{(i)} - \boldsymbol{r}^{(j)}}{\left|\boldsymbol{r}^{(i)} - \boldsymbol{r}^{(j)}\right|^3} \cdot (d\boldsymbol{r}^{(i)} \times d\boldsymbol{r}^{(j)})$$

dove γ_i e γ_j sono due curve orientate chiuse e r(i) è il raggio vettore della prima mentre r(j) è il raggio vettore della seconda. Per curve chiuse G è un numero intero, diverso da zero per curve concatenate fra loro. Il segno di G, la chiralità della concatenazione, dipende dalla possibilità di associare un verso alle curve. Per le proteine il verso va in maniera naturale dall'N-terminale al C-terminale. Essendo però le proteine oggetti tridimensionali discreti è chiaro che occorre passare alle sommatorie su curve discretizzate e imponendo che γ_j sia un loop nel senso definito prima si giunge a

$$G'_c(i,j) \equiv \frac{1}{4\pi} \sum_{i=i_1}^{i_2-1} \sum_{j=j_1}^{j_2-1} \frac{\boldsymbol{R}_i - \boldsymbol{R}_j}{\left|\boldsymbol{R}_i - \boldsymbol{R}_j\right|^3} \cdot (\Delta\boldsymbol{R}_i \times \Delta\boldsymbol{R}_j).$$

dove R(i) = $\frac{1}{2}$(r(i)+r(i+1)) e ΔR(i) = r(i+1)-r(i) ,dove r(i) sono le coordinate degl'atomi dei C_α, R(i) le posizioni intermedie, ΔR i vettori di legame. Massimizzando | G'(i,j) | su tutti i possibili filamenti γ_i, non sovrapposti in sequenza al loop γ_j, si associa quindi un valore di GE ad ogni loop γ_j. Per poter svolgere questo primo calcolo è stato scelto di imporre che il thread così come il loop dovessero essere lunghi almeno 10 amminoacidi.

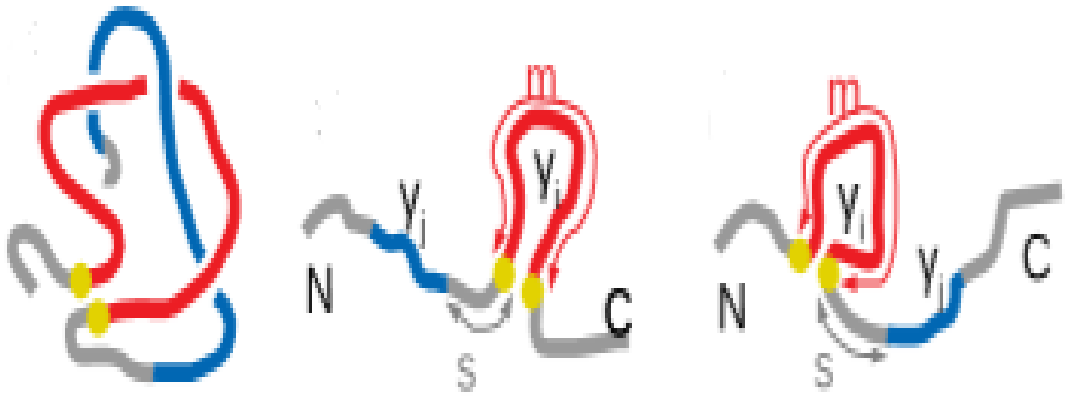


Figura 8. Motivi strutturali topologicamente complessi: in rosso è rappresentato il loop con i due puntini gialli che rappresentano l'inizio e la fine del loop. In blu invece c'è il filamento che a seconda che si trovi a sinistra o a destra del loop viene definito come N-thread o C-thread. La lunghezza del loop è m, s è la distanza tra il loop e il thread.

Capitolo 2

ANALISI TOPOLOGICA

2.1 **Procedura di clustering**

In [1] sono stati calcolati i valori di GE per tutti i loop (circa 3.6 milioni) presenti in un insieme di circa 16000 proteine a singolo dominio [12] con un filtro di omologia al 35%. Spieghiamo ora come associare un "peso statistico" ad ogni loop [1]. Ricordiamo che il GE calcolato per ogni loop lo associa ad uno specifico filamento per cui il corrispondente | G'(i,j) | è massimo.

Una volta esaminata la struttura tridimensionale e riconosciuta la posizione dei singoli amminoacidi lungo la catena (tramite il loro C_α), si evidenzia la necessità di suddividere in cluster i tanti loop della proteina stessa, attribuendo a ciascuno un "peso statistico" dato da 1/N (dove N è il numero di loop contenuti nel singolo cluster).

Per stabilire un criterio che determini quali siano i loop compresi nello stesso cluster si è scelto di considerare come tali quelli che distano mediamente meno di 10 unità amminoacidiche tra loro [1]; in questa definizione si include, in realtà, anche la distanza in sequenza fra i filamenti associati al loop e si considerano inoltre le differenze di GE con un peso opportuno.

L'ipotesi principale è che loop nello stesso cluster siano correlati fra loro e quindi occorra attribuire a tutto il cluster un peso statistico pari ad 1: ovvero, il tutto conta come un singolo elemento indipendente. Cluster diversi si assumono invece come indipendenti e non correlati fra loro.

In formula, detti ($i_{(1)}$; $i_{(2)}$) gli amminoacidi di inizio e fine del loop,($j_{(1)}$;$j_{(2)}$) l'inizio e la fine del filamento per cui si ottiene un entanglement massimo e $G_{(i,j)}$ il GE relativo all'accoppiamento, si ha che, due loop A e B, appartengono allo stesso cluster se

$$d_{AB} = \sqrt{(i_1^A - i_1^B)^2 + (i_2^A - i_2^B)^2 + (j_1^A - j_1^B)^2 + (j_2^A - j_2^B)^2 + w_g[G'_c(i^A, j^A) - G'_c(i^B, j^B)]^2}$$

è minore o uguale a 20 con w_g=10^4. Sebbene le scelte di d_{AB} e di w_g siano arbitrarie, entro loro variazioni ragionevoli i risultati non cambiano significativamente [1]. In questa maniera si può definire il cluster dei loop correlati ad un dato loop, per ogni scelta di quest'ultimo.

I diversi cluster così ottenuti hanno però dei loop in comune ed è quindi necessario scegliere un secondo criterio che permetta di assegnare ogni loop univocamente in un solo cluster. La scelta adottata è quella di inserire i loop in comune (tra due o più cluster) nel cluster più grande.

Questa non è ancora una procedura univocamente definita, perché, a parità di grandezza dei cluster più grandi, i loop vengono inseriti casualmente in uno di essi.

2.2 **Asimmetria e chiralità**

Al termine dell'operazione di clustering, di cui si omettono le tabelle a causa del numero molto elevato di loop (come già detto, sono circa 3.5 milioni totali in più di 16 mila proteine analizzate), si ha a disposizione il voluto "peso statistico" da attribuire a ciascun loop. Siamo interessati in particolare ad evidenziare asimmetrie statisticamente significative nei conteggi di N- e C-thread al variare del corrispondente GE.

La presenza di asimmetrie potrebbe essere legata all'importanza nel processo evolutivo del ripiegamento cotraslazionale, quando le proteine iniziano a ripiegare dal lato N-terminale mentre vengono sintetizzate nel ribosoma. Quantifichiamo l'asimmetria fra N- e C-thread usando il parametro $Nd = \frac{Nn-Nc}{Nn+Nc}$ dove Nn è il numero di N-thread e Nc è il numero di C-thread.

Usando, allora, il valore di G′ (i,j) per ogni loop, un istogramma Nd($G'_{(i,j)}$), per semplicità indicato con Nd(G′), ricavato con bin di larghezza ΔG=0.05, permette di individuare particolari intervalli di GE in cui è possibile trovare asimmetrie statisticamente significative.

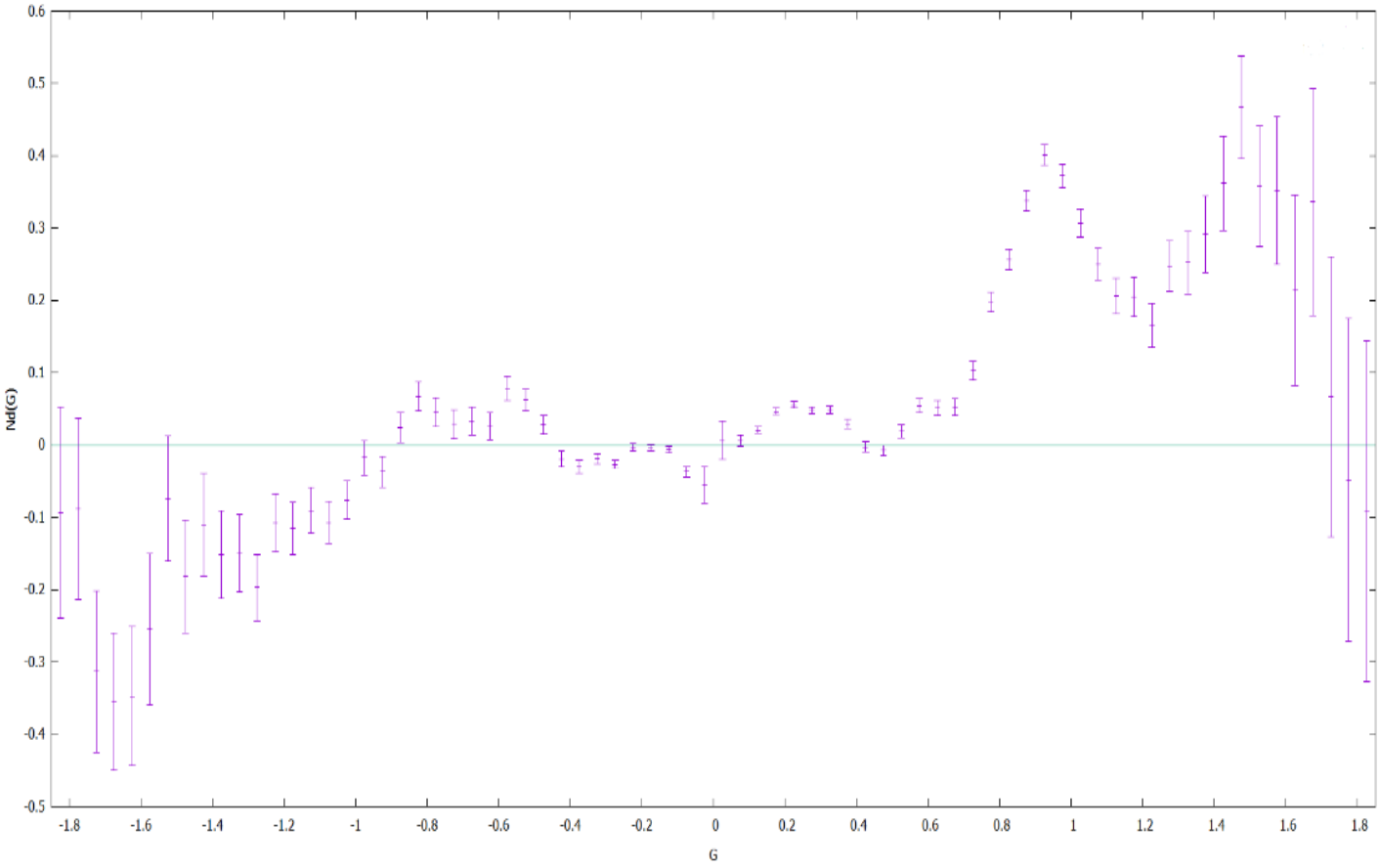


Figura 9: gli Nd sono le differenze tra gli N-thread e i C-thread normalizzate rispetto ai loop totali.

Gli errori nel grafico precedente e nei successivi sono indicati come barre verticali e sono dati dalla formula di propagazione degli errori [6] usando una statistica poissoniana per conteggi dei loop.

Dal grafico precedente è evidente che per $g \geq 0$ (chiralità positive) c'è una netta asimmetria che favorisce gli N-thread che ha i suoi picchi per $0.75<G'<1.2$ e per $1.25<G'<1.7$ mentre per chiralità negative ($g \leq 0$) si ha un solo intervallo di asimmetria per $-1.85<G'<-1$ che invece favorisce i C-thread ($Nd<0$) in maniera lievemente significativa.

Limitando l'interesse a questi "picchi", si può studiare come l'asimmetria rilevata dipenda dalla lunghezza della proteina L. Per evitare problemi dovuti alla bassa statistica per proteine corte, abbiamo scelto di graficare N_d per le proteine di lunghezza $L \leq l$, al variare di l.

Il grafico mostra una caratteristica molto interessante: per $0.75<G'<1.2$ esiste un valore di soglia per $l \approx 100$ oltrepassato il quale si evidenzia una netta asimmetria. Per gli intervalli $1.25<G'<1.7$ e $-1.85<G'<-1$, invece, finché c'è statistica sufficiente, non si osservano variazioni altrettanto nette.

In generale, interpretare questi segnali non è banale. Infatti [7] è probabile che queste differenze siano dovute alle particolari proprietà strutturali del tunnel ribosomiale e ai diversi vincoli posti da queste al variare della chiralità e della lunghezza della proteina.

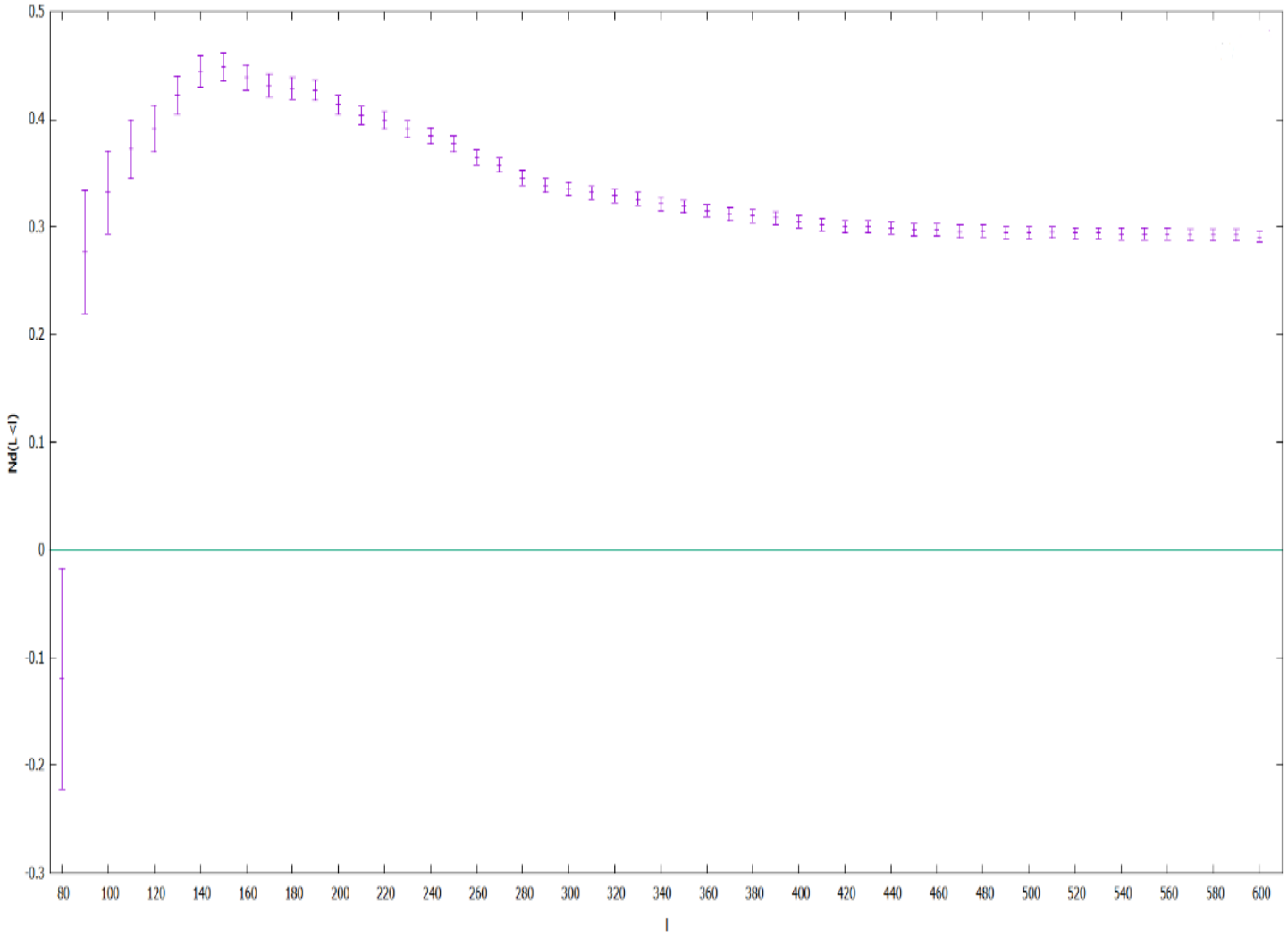


Figura 10:plot relativo all'intervallo $0.75<G'<1.2$ in cui vi è una di soglia per $l\approx 100$ al di sopra della quale si ha prevalenza di N-thread

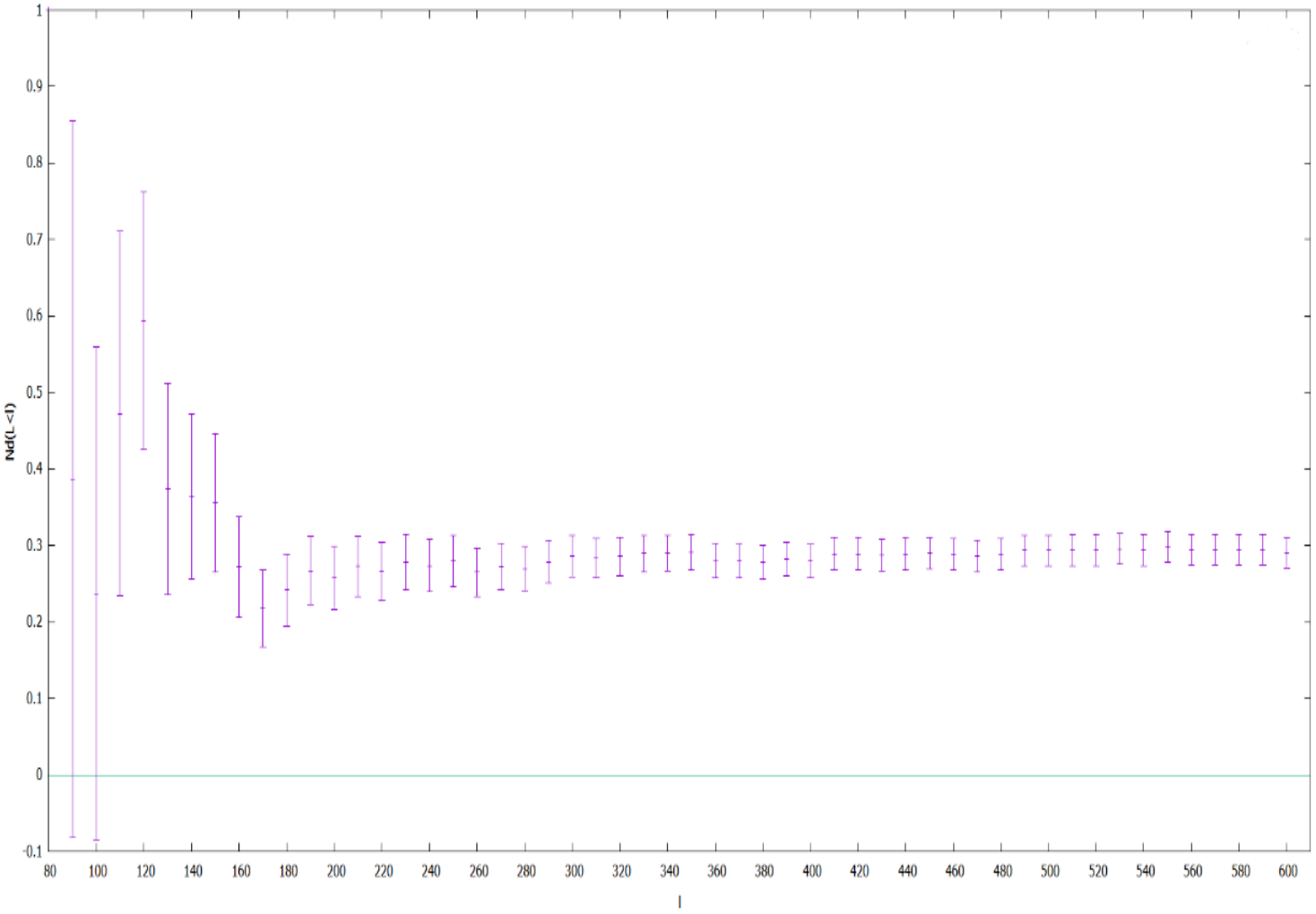


Figura11: plot relativo a 1.25<G'<1.7 in cui scompare la soglia osservata in figura 10 e vi è sempre una prevalenza di N-thread.

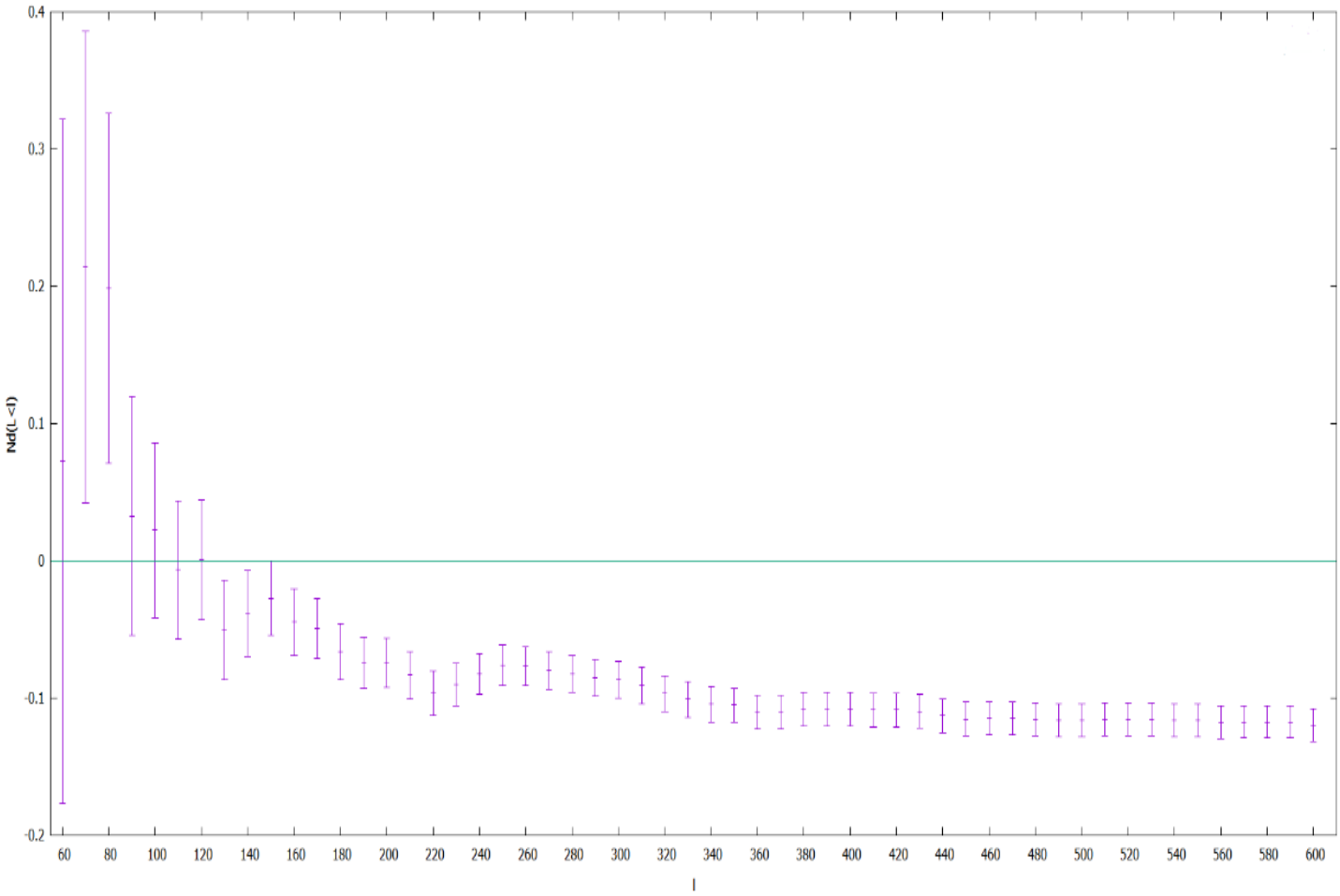


Figura 12: plot relativo a -1.85<G'<-1,finchè l'errore è basso si osserva una prevalenza di C-thread (Nd <0).

2.3 **Distribuzioni delle lunghezze dei loop**

Abbiamo infine analizzato gli istogrammi normalizzati p(m) in funzione della lunghezza m dei loop, per N-thread e C-thread separatamente. Se per chiralità negative ($g \leq 0$) non è apprezzabile alcun tipo di differenza tra le due distribuzioni, lo stesso non accade per chiralità positive ($g \geq 0$). Sovrapposto all'andamento decrescente normalmente atteso per un polimero (per un random walk $p(m) \approx m^{-\frac{1}{2}}$) si nota, invece, solo per i C-thread, un picco piccolo, ma statisticamente significativo per $20 \leq m \leq 30$. Tale picco è noto in letteratura come "picco di Trifonov" e in base a diversi studi ([10]-[11]), dovrebbe essere una prova di come le proteine si siano evolute, partendo da quelle più semplici fino ad arrivare a quelle più complesse. E' molto interessante che la nostra analisi leghi il picco di Trifonov unicamente ai C-thread con chiralità positive; stabilendo quindi un possibile legame tra i meccanismi evolutivi postulati per la sua origine e le proprietà del ripiegamento cotraslazionale nel tunnel ribosomiale.

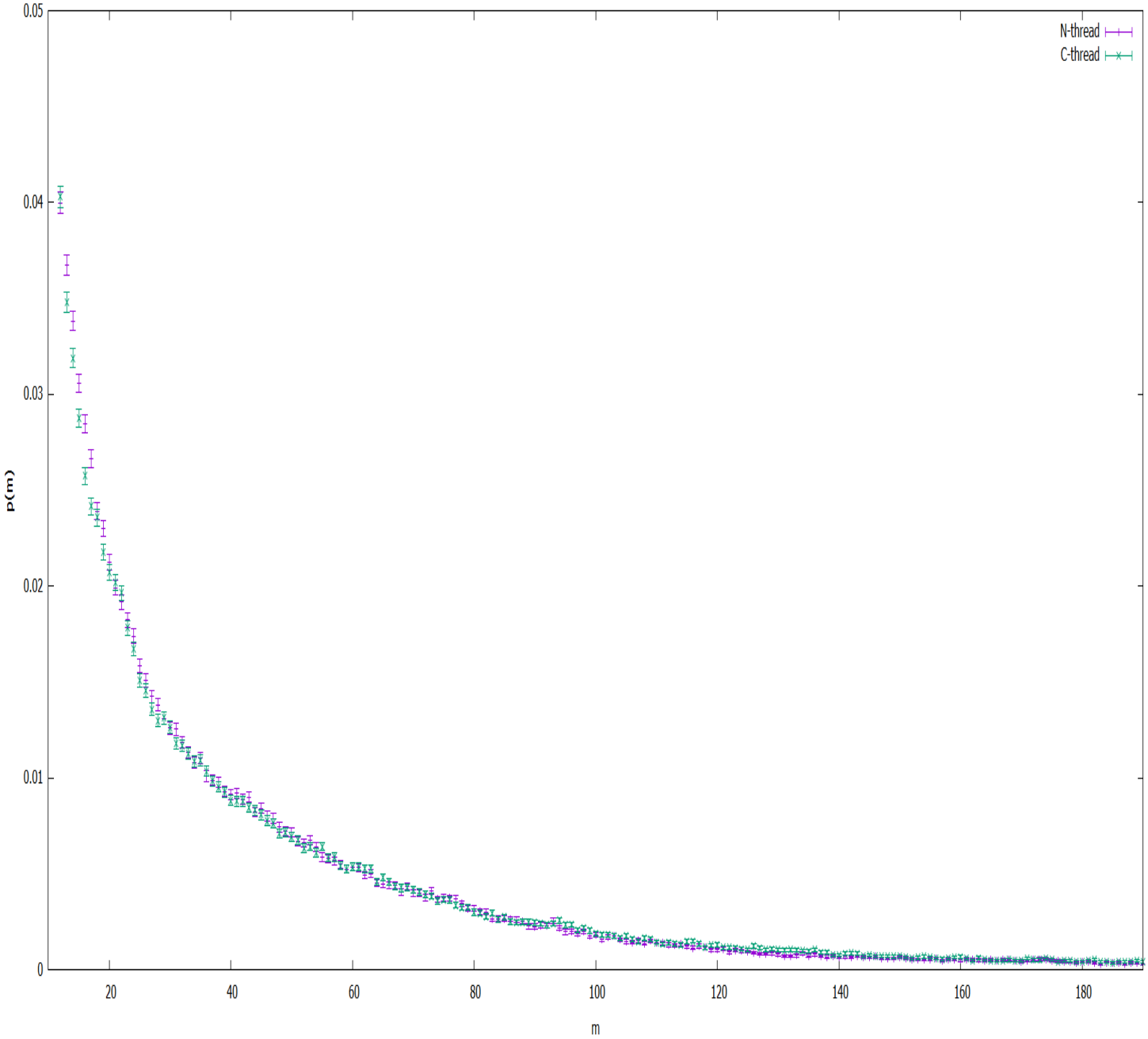


Figura 13.Plot relativo a g≤0.

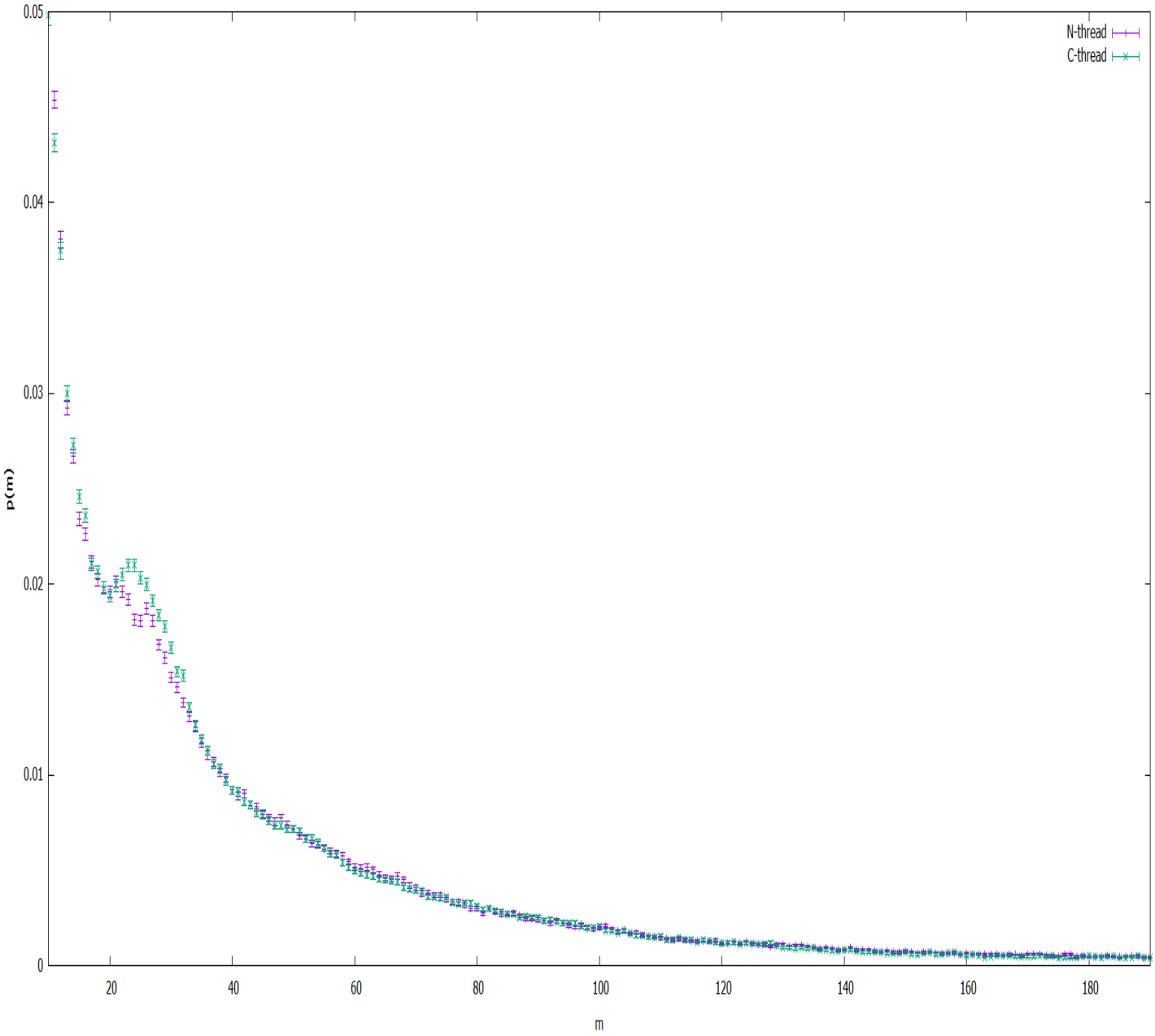


Figura 14. “Picco di Trifonov”([10],[11]) per lunghezze di loop comprese tra 20 e 30 amminoacidi.

Un’analisi simile alla precedente, effettuata per loop “entangled”, ossia caratterizzati da $|G'| \geq 0.8$, non ha mostrato segnali interessanti di differenze tra gli N-thread e i C-thread, probabilmente anche a causa della statistica troppo bassa.

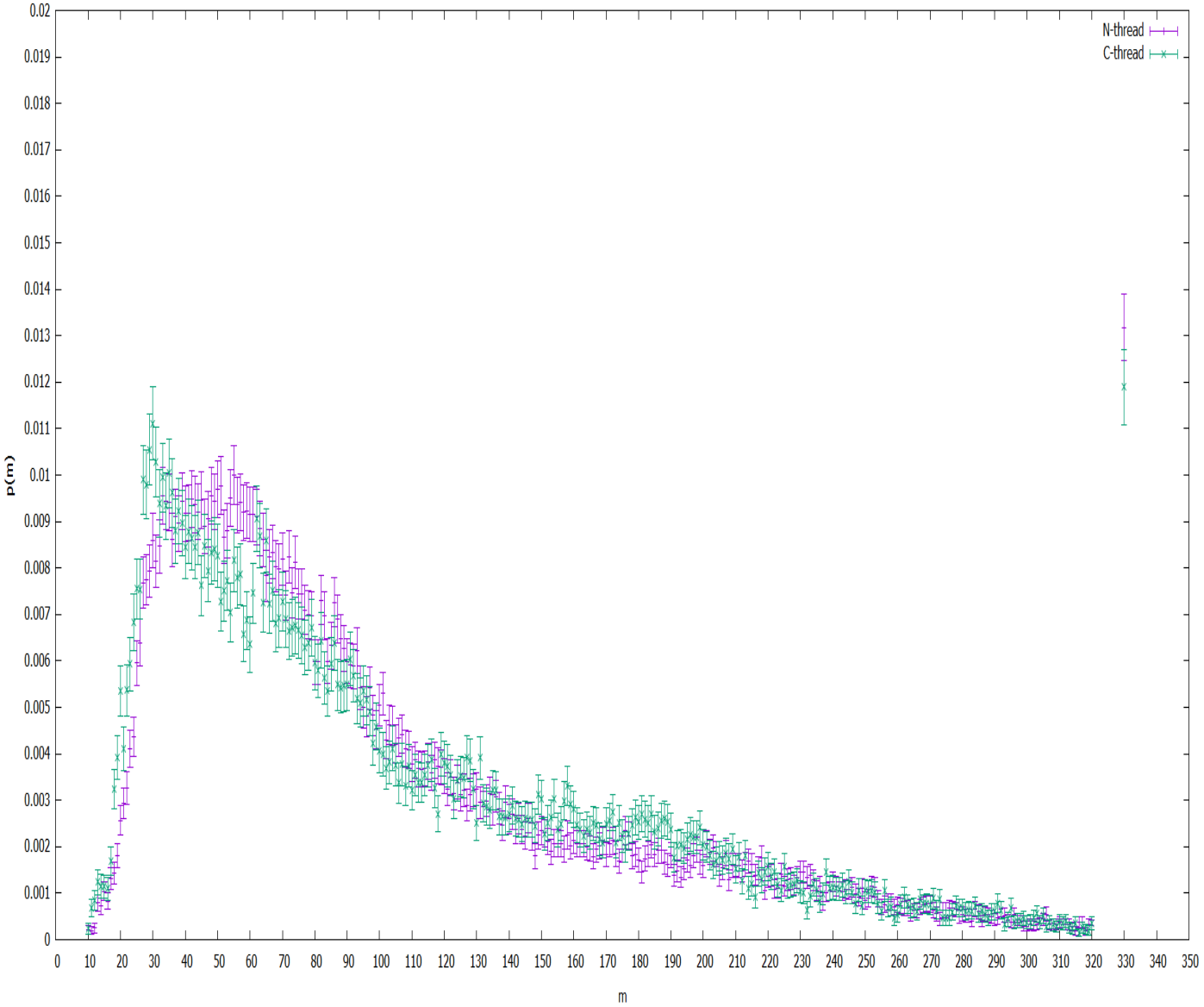


Figura 15: grafico per loop “entangled” , con $|G| \geq 0.8$

CONCLUSIONI E PROSPETTIVE

In questo lavoro di tesi sono stati evidenziati una serie di segnali statisticamente significativi riguardanti i conteggi di motivi strutturali caratterizzati da complessità topologica e rilevati nelle strutture native di proteine. In particolare, ci siamo focalizzati sui conteggi di N-thread e C-thread, termini con i quali denominiamo coppie loop-filamento in cui il filamento "attraversa" il loop e distinti a seconda che il filamento sia più vicino rispetto all' N-terminale (N-thread) o al C-terminale (C-thread). E' stato proposto [1] che un'asimmetria in favore di N-thread possa essere legata all'esigenza di non rallentare il ripiegamento cotraslazionale che avviene mentre la proteina viene sintetizzata nel ribosoma. In quest'ipotesi è più conveniente ripiegare per prima cosa il filamento nella sua struttura nativa e solo dopo avvolgere il loop (che viene sintetizzato dopo il filamento nel caso di N-thread) attorno ad esso. Questa ipotesi viene confermata dai nostri risultati per chiralità positive, mentre per chiralità negative si osserva una lieve ma significativa prevalenza di C-thread (figura 9). Inoltre, l'asimmetria rilevata dipende dalla lunghezza delle proteine considerate in maniera non banale (figure 10,11,12). In particolare, l'asimmetria in favore degli N-thread non è più presente per proteine più corte di 100 amminoacidi nel caso di GE≈ 1. Infine, le distribuzioni per le lunghezze dei loop mostrano come il picco di Trifonov, ben noto in generale [10], si presenti solo nel caso dei C-thread con chiralità positive. Un'interpretazione approfondita dei risultati ottenuti in questa tesi è non banale; infatti, per comprendere a fondo il folding cotraslazionale occorre conoscere in maniera approfondita ciò che accade all'interno del tunnel ribosomiale di una cellula vivente.

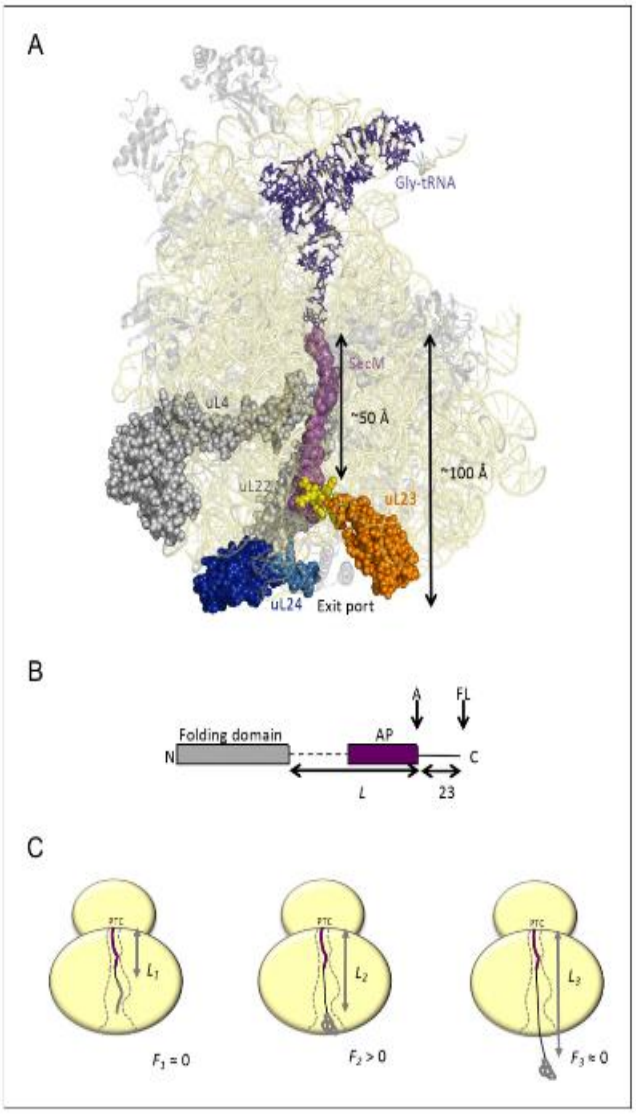


Figura16.Questa immagine [9], per esempio, è una buona illustrazione che ben evidenzia le difficoltà legate ad uno studio dettagliato del folding cotraslazionale. Concentrandoci solo sulla parte A e B possiamo dire che nella prima c'è una visione frontale della subunità ribosomiale dell'E.coli(un batterio) con due tunnel proteici uL4 e uL22 indicati in grigio. Il dominio globulare uL23 è indicato in arancione con il loop in giallo. uL24, un altro tunnel ribosomiale, è mostrato in blu scuro, con il loop all'uscita del tunnel mostrato in blu chiaro. L'uscita del tunnel, di circa 100 *Å* in lunghezza, è sottolineata da una catena nascente di color viola. Nella parte B invece, si rappresenta l'arresto del peptide costituito da L amminoacidi. Un calcolo relativo alla forza necessaria per l'interruzione della produzione della catena fornisce $f_{FL} = I_{FL}/(I_A+I_{FL})$ dove I_A e I_{FL} sono l'intensità delle "bande" di arresto che corrispondono ai prodotti A e FL.

Ad oggi questo argomento viene studiato molto attivamente e ci sono evidenze, sia sperimentali che numeriche, che proteine anche non piccolissime (100 residui) possano ripiegare o iniziare a ripiegare mentre sono ancora confinate nel tunnel ribosomiale.

E' quindi ragionevole ipotizzare che la particolare forma tridimensionale del tunnel ribosomiale, possa causare la relazione fra chiralita' e asimmetria dei motivi entangled identificata nei nostri risultati. Ovviamente serviranno altri studi, sia numerici che sperimentali,per comprovare questa ipotesi.

REFERENZE:

[1] M. Baiesi, E.Orlandini , F.Seno, A. Trovato(2019). Sequence and structural patterns detected in entangled proteins revel the importance of cotranslational folding. “Materials and Methods”.ScientificReports 9:8426.

[2] Kessel ,Ben-Tal(2008).Introduction to proteins.Structure,function and motion.Second Edition.Chapter 3. NewYork: CRC Press Taylor &Francis Group.

[3] Kessel ,Ben-Tal(2008).Introduction to proteins.Structure,function and motion.Second Edition. Chapter 2. NewYork: CRC Press Taylor &Francis Group.

[4]Finkelstein,Ptitsyn(2002).Protein-Physics-Acourse of Lectures.Lecture1. London:Academic Press

[5]Finkelstein,Ptitsyn(2002).Protein-Physics-A course of Lectures.Lecture1-2. London:Academic Press

[6]M. Loreti(2006).Teoria degli errori e fondamenti di statistica.Introduzione alla fisica sperimentale. Capitolo 10.Dipartimento di Fisica. Università di Padova.Libro composto e stampato dall’autore http://wwwcdf.pd.infn.it/labo/book.pdf

[7]O'Brien,F.Trovato(2016).Insights into cotraslational nascent protein behavoir from computer simulations.Annu Rev Biophys.45:345-69

[8]G.von Heijne,B.Nilsson,R. Hedman,J.Marino(2015).Cotraslational protein folding inside the ribosome exit tunnel. Cell Reports 12,1533-1540

[9]R.Kudva,P.Tian,F.Pardo-Avila,M.Carroni, R.Best,H.Bernstein,G. von Heijine (2018) The shape of the bacterial ribosome exit tunnel affects cotranslational protein folding. eLife ;7:e36326

[10]N.Berezovsky,A.Grosberg,N.Trifonov (1999).Closed loops of nearly standard size: common basic element of protein structure. FEBS Letters 466(2000) 283-286

[11]E.Trifonov, I. Berezovsky(2003). Evolutionary aspects of protein structure and folding.Current Opinion in Structural Biology. 13:110-114

[12] Dawson, N. et al. Cath: An expanded resource to predict protein function through structure and sequence. Nucleic Acids Research 45, D289–D295 (2017).